NASA Crew Interface Measurement: Phase II

Ian Robertson, KBR
Kritina (Tina) Holden, Leidos
Michael Bishop, KBR
Tyler Duke, Leidos

2022 NASA Human Research Program Investigators' Workshop

Background

- In NASA programs, spacecraft developers must formally verify their products in order to be certified for flight
  - Designs must meet predetermined success criteria in order to “pass”
- Verification testing involves astronauts performing realistic tasks with flight-like equipment in a realistic simulation, and collection of human-system performance data including:
  - Usability (System Usability Scale)
  - Workload (Bedford Workload Scale)
  - Errors
  - Acceptability (Astronaut Office Acceptability Scale)
  - Subjective comments

Project Motivation

The NASA Commercial Crew Program (CCP) recently completed verification testing.

Crew feedback during CCP verification testing indicated areas for improvement:

- Test time should be reduced, if possible
- Scale phrasing should be better tailored for the NASA domain

Project Goals

Improve verification testing for future spaceflight programs by reducing test time where possible, and better tailoring measurement tools for NASA by:

- Identifying if any measures employed in verification testing are redundant
- Tailoring the System Usability Scale for use at NASA
  - Identify changes that better suit the NASA domain
  - Formally assess the modified scale via psychometric methods

CCP Data Analysis

Analysis of CCP Data

- A psychometric analysis of the CCP data was just completed
- The analysis had the following goals:
  - Qualify the relationship between the different verification measures (e.g., correlation)
  - Identify the unique, and overlapping, contributions each measure makes to verification decisions
- Initial results are discussed here

Analysis of CCP Data

Verification data were obtained from two providers (e.g., Boeing, SpaceX)

- *N* = 26 participants across the 2 providers in the analysis
- Some crew participated in testing for both providers
- Each provider designed and organized verification tests in their own way

Data Analysis

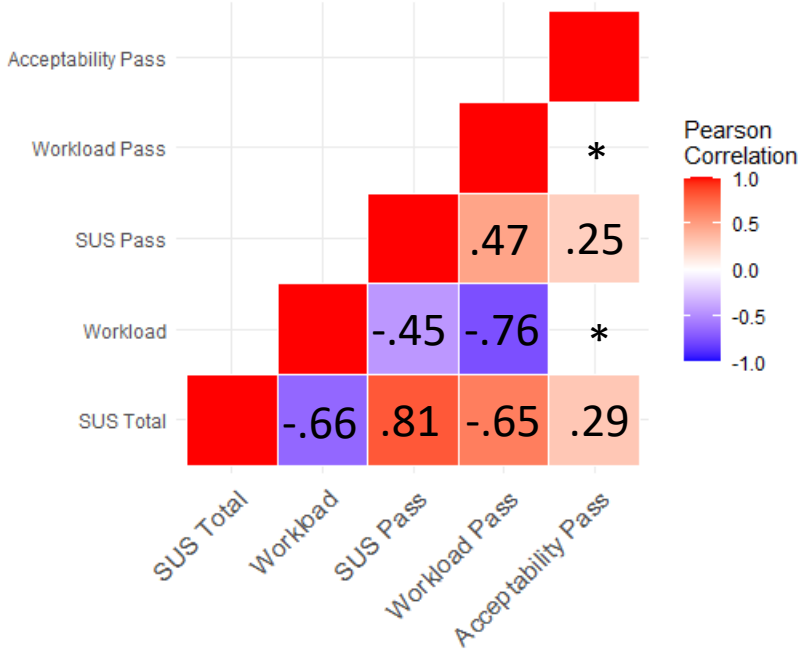
- The relationships between measures were analyzed across two tasks
  - Cockpit (e.g., computer-based tasks)
  - Physical (e.g., replacing filters, egress, cabin reconfiguration)

- The relationship between measures was assessed with Pearson's correlation
- Agreement between measures was assessed with Fleiss's Kappa

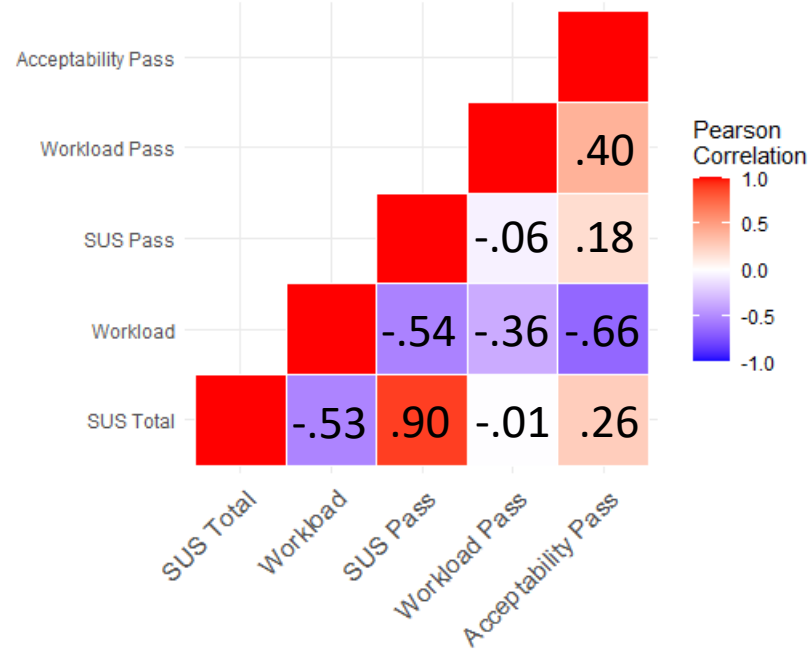
Correlations Between Measures

Measure	Scale
System Usability Scale	0-100
Bedford Workload	1-10
SUS Pass	Pass = 85+
Workload Pass	Off nominal Pass = < 6 Nominal Pass = < 3
Acceptability	Pass = “Acceptable”

Physical



Cockpit



* = Insufficient data

Verification Agreement

Fleiss' Kappa (0 = no agreement; 1 = high agreement) was used to assess overall agreement between the verification decisions for each measure (SUS, Workload, and Acceptability)

- Physical: Fleiss' $\kappa = .073$, $p = .50$
- Cockpit: Fleiss' $\kappa = .29$, $p < .001$
- **May** indicate low overall agreement between the three verification outcomes

Possible reasons:

- Each measure may be capturing different aspects of the verification decision
- Invariance-ICCs produce lower values when data is invariant
- Currently no final conclusions have been reached

CCP Data Analysis Summary

Observed relationships make theoretical sense and measures are not redundant

- Usability is inversely related to workload
- Observed positive association between usability and acceptance ratings
- Observed associations may indicate that measures are not redundant

Fleiss' κ indicates little agreement between verification decisions between the three measures

- Practically, may indicate that each measure is useful
- Could be due to other reasons (yet to be explored)

Final analysis is still ongoing

Development of the NASA Modified SUS

Prior Work: SUS Modifications

- It was clear that the phrasing in several statements could be modified to better fit the NASA domain
- It has been demonstrated in the literature that SUS is still valid with small changes (e.g., phrasing or removal of an item)
- The goal was to make minimal changes that would reduce negative comments that could be impacting ratings and produce results equivalent to the original SUS
- After several review and discussion sessions, two statements were deleted, and three statements were rephrased to create the NASA Modified SUS (NMSUS)

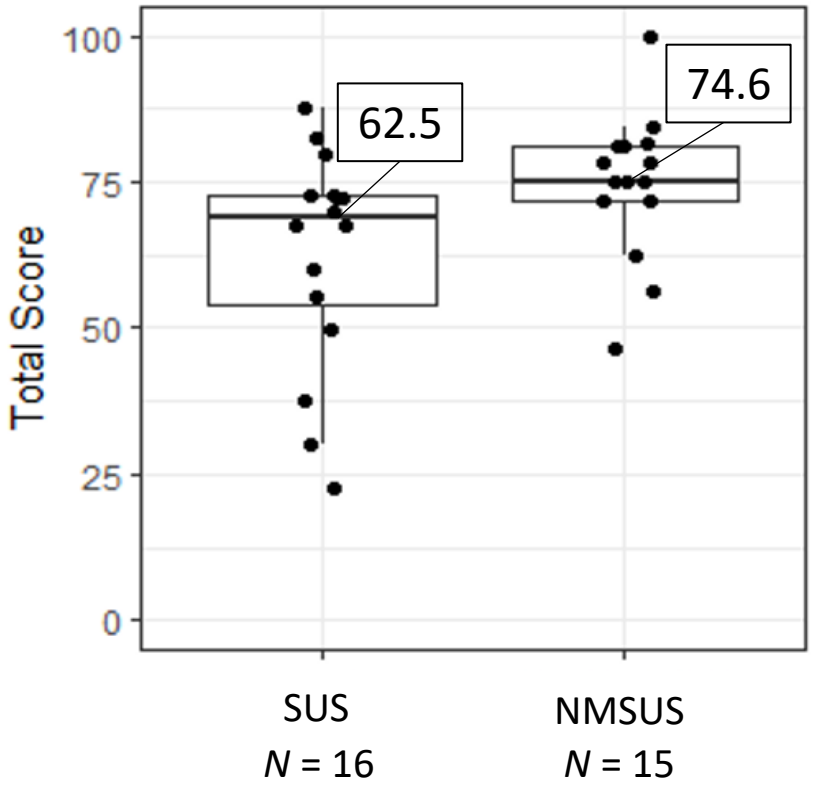
Prior Work

- Pilot tested the modified SUS with crew-like participants in preparation for full validation

Results

- NMSUS was reliable Cronbach's α = .77 [.60, .93]
- Observed difference of *M* = 12.1 points (*p* < .05)

- Pointed to need to perform additional reliability and validation analysis



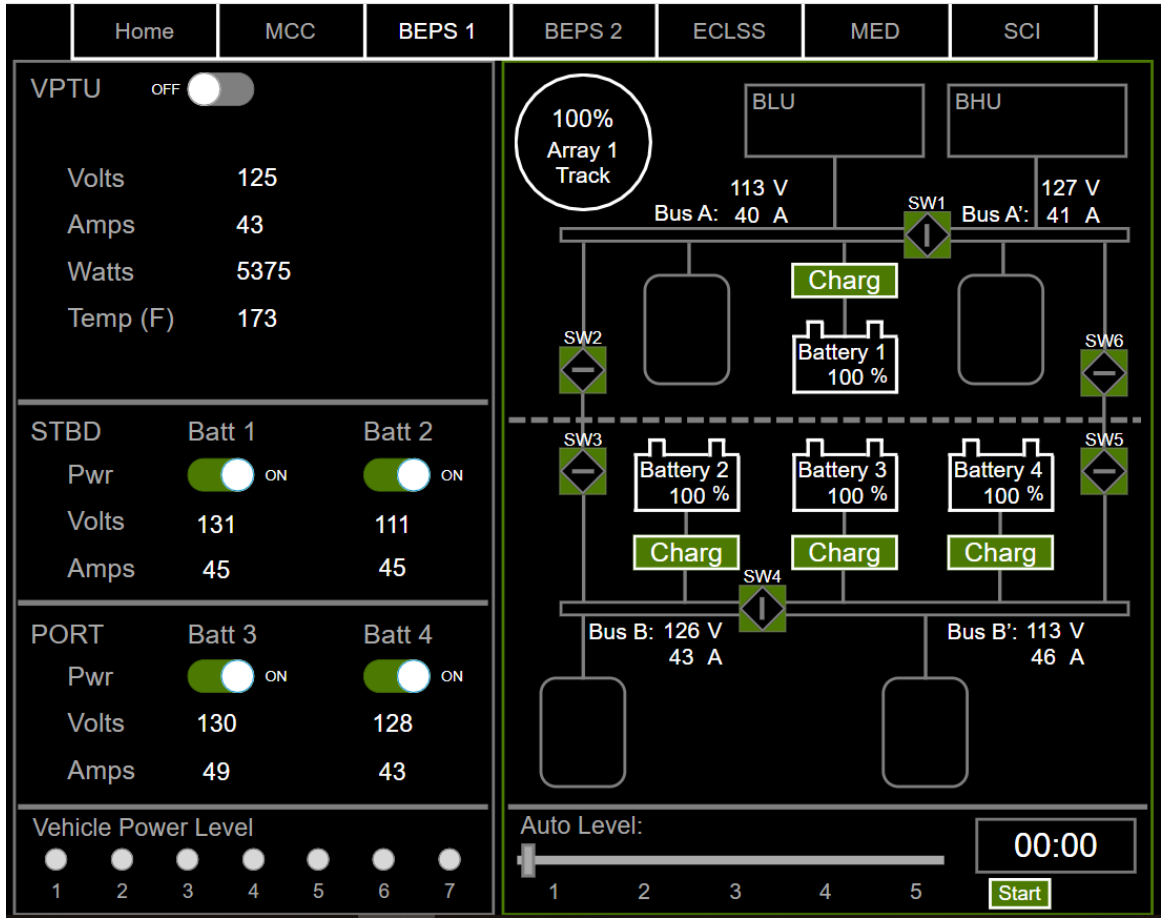
Current Study

The goals of the current study are:

- To determine if the observed difference between the SUS and NMSUS are due to changes of the psychometric properties of the scale or some other reason (e.g., sampling variance)
- Conduct additional reliability analyses on the NMSUS
- Collect evidence for the validity of the NMSUS (revalidation necessary do to changes to wording and dropping items)

Validation and Reliability Study: Method

- Participants will complete two representative tasks with representative prototype vehicle interfaces
- Participants will rate both interfaces using the SUS and/or the NMSUS
- Additional usability data will be collected such as number clicks, errors, time on task, and user comments



Example of interfaces used in study

Validation and Reliability Study: Analyses

Question 1: Do the results of Study 1 represent a true difference in scale scores between the two versions of the SUS?

- Conduct correlations between the SUS and NMSUS for both interfaces
- Conduct t-test between SUS and NMSUS scale scores for both interfaces

Question 2: Is the NMSUS reliable?

- Perform classic reliability analyses of NMSUS (Cronbach's alpha)
- Did rewording items change how participants respond to those items (correlations between original and reworded items for both interfaces)?

Validation and Reliability Study: Analyses

Question 3: Is the NMSUS valid?

- Assess convergent validity by correlating NMSUS with other measures of usability (e.g., other subjective measures)
- Assess criterion validity: How do NMSUS scale scores relate to other usability outcomes (e.g., errors and time on task) Are NMSUS scores able to predict participants' preference for one interface over the other?
- Qualitatively compare performance of NMSUS to SUS on the same metrics

NMSUS Summary

In prior work modifications were made to the System Usability Scale to better suit NASA users

- Results of the first study indicated that there may be differences in scale scores between the SUS and NMSUS

Current work expands upon the initial study by collecting more data in order to:

- Address if difference observed in the first study reflect true scale score differences or something else (likely sampling variance)
- Perform additional reliability analyses on the NMSUS
- Collect evidence regarding the validity of the NMSUS